Optimal Linear Joint Source-Channel Coding with Delay Constraint

Erik Johannesson, Anders Rantzer, Fellow, IEEE, Bo Bernhardsson, and Andrey Ghulchak

Abstract—The problem of joint source-channel coding is considered for a stationary remote (noisy) Gaussian source and a Gaussian channel. The encoder and decoder are assumed to be causal and their combined operations are subject to a delay constraint. It is shown that, under the mean-square error distortion metric, an optimal encoder-decoder pair from the linear and time-invariant (LTI) class can be found by minimization of a convex functional and a spectral factorization. The functional to be minimized is the sum of the well-known cost in a corresponding Wiener filter problem and a new term, which is induced by the channel noise and whose coefficient is the inverse of the channel's signal-to-noise ratio. This result is shown to also hold in the case of vector-valued signals, assuming parallel additive white Gaussian noise channels. It is also shown that optimal LTI encoders and decoders generally require infinite memory, which implies that approximations are necessary. A numerical example is provided, which compares the performance to the lower bound provided by rate-distortion theory.

Index Terms—Analog transmission, causal coding, delay constraint, joint source-channel coding, MSE distortion, remote source, signal-to-noise ratio (SNR).

I. Introduction

THE design of systems for point-to-point communication of analog data over noisy communication channels has a theoretical basis in Shannon's separation theorem. The theorem gives a bound on the optimal performance theoretically achievable (OPTA) by any communication system. Specifically, it says that the distortion can not be made smaller than $\mathcal{D}_{\min}$, which can be obtained from

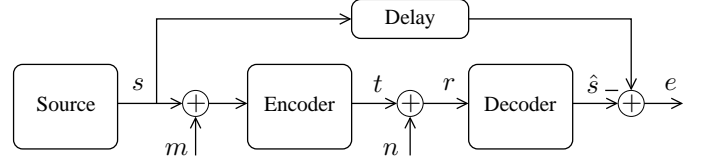$$\mathcal{R}(\mathcal{D}_{\min}) = \mathcal{C}, \qquad (1)$$

Fig. 1. The encoder measures the source sequence s plus the measurement noise m and transmits t over the channel. The decoder receives t plus the channel noise n and forms $\hat{s}$, the estimate of s. Each source element has to be estimated after a given delay in order to minimize the error e.

where $\mathcal{R}(\mathcal{D})$ is the rate-distortion function, which is given by the source statistics and the distortion measure, and $\mathcal{C}$ is the channel capacity. Under appropriate assumptions, the separation theorem also shows that it is possible to come arbitrarily close to $\mathcal{D}_{\min}$ by the combination of source coding and channel coding. These codes can, in principle, be independently developed without loss. This means that the channel code designer does not need to know anything about the source, and vice-versa, which is clearly a practical advantage.

The separation theorem does, however, rely on asymptotic arguments where the delay and the size of the codebook are allowed to increase indefinitely. Consequently, it does not hold in presence of delay or complexity constraints and imposing such constraints generally renders the distortion bound unachievable. Since infinitely large delays or codebooks are not possible in practice, a suboptimal performance may have to be accepted. Moreover, to minimize the distortion in the presence of these constraints, it may be necessary to abandon the separation-based design and consider joint source-channel codes.

This is the subject of the present paper, where we consider transmission of a stationary colored Gaussian source over a power-constrained channel with additive colored Gaussian noise, under the mean-square error (MSE) distortion criterion. The encoder and decoder are constrained to be causal and their combined operations are subject to a delay constraint. Further, we allow for the possibility of a remote (noisy) source. The situation is illustrated in Fig. 1.

The encoder and decoder will be restricted to the class of linear and time-invariant (LTI) filters. The linearity assumption and the additive noise models allow us to formulate the distortion minimization as a transfer function optimization problem. The main result is that a jointly

optimal encoder-decoder pair from the LTI class can be found by first minimizing a functional of the form

$$\|R - X\|_2^2 + \frac{1}{\sigma^2} \|XN\|_1^2, \qquad (2)$$

where $R \in \mathcal{L}_\infty$ and $N \in \mathcal{H}_\infty$ are given transfer functions and σ^2 is the signal-to-noise ratio (SNR), over $X \in \mathcal{H}_2$. The encoder and decoder are then obtained from a spectral factorization. A corresponding result is also shown to hold in the case with vector-valued signals and parallel additive white Gaussian noise (AWGN) channels.

The restriction to linear encoders and decoders may obviously result in suboptimal solutions. Nevertheless, the linear solution to any problem instance will provide an upper bound to the minimum distortion possible for the given SNR, delay constraint, and signal spectra. Moreover, the proposed design methods are relatively simple and computationally feasible.

An application where this problem formulation could be relevant is the transmission of speech in mobile communication. The source signal to be estimated at the receiver is the speech signal. The delay constraint is based on the acceptable latency and the noise is any background sound present at the microphone.

The rest of this section will present the relevant previous research and alternative interpretations of the problem. Section II presents the mathematical notation used in this paper. The exact problem formulation is given in Section III. Section IV is devoted to the solution of the problem, first in the scalar and then in the vector case, followed by a theorem stating that optimal LTI encoders and decoders require infinite memory. Section V presents a procedure for numerical solution and a numerical example where the performance of the optimal LTI encoders and decoders is compared to the lower bound provided by the separation theorem. Finally, Section VI presents the conclusions and discusses further research. Some technical lemmas have been put in the appendix.

A. Previous Research

The problem studied in this paper is closely related to that of finding the optimal modulation matrices for linear coding and decoding of a Gaussian vector source for transmission over a Gaussian vector channel. Optimal modulation matrices were derived in [2], where it was also shown that linear modulation is only optimal when the source and channel can be matched. That is, when their dimensions match and the source and channel noise covariance matrices can be diagonalized into uniform variances. The same problem was considered in [3], where the solution was also given for the case when the channel components have individual power constraints. The performance of optimal linear coding was compared, for a number of cases, to the OPTA, given by (1), in [4].[1]

The general suboptimality of linear coding arises from the fact that it cannot match any colored Gaussian source to any colored Gaussian channel. It has recently been shown, however, that such matching can be achieved by the combination of prediction and modulo-lattice operations [5].

The problem of coding with a remote source was first considered for the Gaussian case with additive noise and MSE distortion in [6]. It was shown that the problem is asymptotically equivalent to, and can thus be reduced to, the fully observed case and that an optimal encoder generally has a structure consisting of an optimal estimator followed by optimal encoding for a noise-free source. This structural result was generalized to the non-gaussian and finite time horizon cases in [7]. The problem was further studied in [8], where it was noted that in the case of white source noise, the criterion in the reduced problem is given by the conditional expectation of the original criterion given the encoder input. It was pointed out in [9] that the equivalence in [6] actually was proved for the one-shot problem as well. Moreover, it was shown that the reduction to the non-remote problem follows from a general "disconnection principle". In the literature, the problem of coding with a remote source often includes the possibility of noise at the receiver as well. The main motivation for excluding that possibility here is the fact, noted in [7], that the optimality of an encoder-decoder design is independent of additive and independent zero-mean noise at the receiver.

Coding problems with delay constraints have not received the same level of attention as their classical counterparts. Some structural results have, however, been obtained. The optimal causal source coder for a white source has been found to be memoryless [10]. For a Markov source of order k and delay constraint d, an optimal real-time source coder only needs to use the last $\max\{k, d+1\}$ source symbols plus the current state of the decoder. No such memory bound is given, however, when the encoder does not have access to the decoder state [11]. Joint source-channel coding with noiseless feedback was considered for finite alphabet sources in [12] where it was demonstrated that feedback is useful in general, but that coding is useless for a class of channels with a certain symmetry property. The results in [11], [12] have been generalized in [13], which also gives a nice overview of the literature on real-time coding. Conditions have also been found for when optimal performance can be achieved without coding (even when allowing coding systems with arbitrary delay) [14].

Since the OPTA given by (1) cannot generally be achieved in the presence of delay constraints, a relevant question to ask is of course what the OPTA is when there

are such constraints. A partial answer in the form of upper bounds on the rate-distortion functions for zero-delay and causal source coding is given in the recent paper [15]. Interestingly, some of the results in that paper are obtained by solving a problem which is somewhat similar to the one considered in this paper. The solution of that problem can be applied to solve some particular instances of the problem considered in this paper. The main difference is that they assume that the encoder has access to noiseless feedback from the channel output. Moreover, only the scalar case with zero delay constraint and no noise at the source is considered. The same problem has previously been considered in [16], [17] as a means to design optimal scalar feedback quantization schemes.

Real-time source coding for a remote source has been considered in [18]. The structural results of [11], [12] were extended to cover remote sources in [19], which also presented a separation result for the linear-quadratic Gaussian case similar to the one in [6]. A method for design of optimal real-time coding systems for noisy channels was presented in [20] using noisy feedback and in [21] without feedback. However, there seems to be no method for efficient numerical application of the solution.

B. Alternative Interpretations

It is possible to make two alternative interpretations of the problem illustrated in Fig. 1.

1) Connection to Wiener Filter: The problem of estimating a signal that is measured with additive noise under an MSE criterion is solved by the Wiener filter [22]. The filter is usually obtained by solving the Wiener-Hopf equations, but can also be expressed in the frequency domain as the stable filter K that minimizes

$$\left\|(z^{-d} - K)S\right\|_2^2 + \|KM\|_2^2, \tag{3}$$

where d is the allowed time delay and S and M are transfer functions that represent the frequency characteristics of the signal of interest and the measurement noise, respectively.

It is possible to interpret the problem in Fig. 1 as a distributed Wiener filtering problem, where the filter is separated into two different locations. The communication channel is used to model the communication constraint between the two locations. This interpretation is strengthened by the fact that minimization of (3) is equivalent to minimizing

$$\|R - X\|_2^2, \tag{4}$$

where R is the same transfer function as in (2), over $X \in \mathcal{H}_2$. Comparing (4) with (2) it is seen that the cost in the present problem is equal to the cost in a Wiener filtering problem plus an additional term, which is induced by the communication channel. Since the coefficient of the new term is the inverse of the channel's SNR, the cost is asymptotically equal to that in the Wiener filtering problem when the SNR tends to infinity.

2) As a Feed-Forward Control Problem: Fig. 1 may be interpreted as follows: The source signal is a disturbance that will affect some system where a controller (the decoder) can compensate. The controller has a remote sensor that measures the disturbance and transmits information to the controller over the channel. In this interpretation the delay block may also include any dynamics that the disturbance passes through on the way. A similar interpretation was discussed in [9].

A similar problem setup was studied in [23], where information theory was used to find a lower bound on the reduction of entropy rate made possible by side information communicated through a general channel with known capacity. Under stationarity assumptions, this was used to derive a lower bound, which is a generalization of Bode's integral equation, on a sensitivity-like function.

II. Notation

The techniques in this paper rely on concepts from functional analysis, such as $\mathcal{L}_p$ (Lebesgue), $\mathcal{H}_p$ (Hardy) and $\mathcal{N}^+$ (Smirnov) function classes and inner-outer factorizations. To conserve space, only some of the most important facts will be given here. The interested reader is referred either to [1] or to [24], [25] and [26] for the remaining relevant definitions and theorems.

The natural logarithm is denoted log. The complex unit circle is denoted by $\mathbb{T}$. The singular value decomposition of A is taken as $A = U\Sigma V^*$, where Σ is square. A singular value decomposition of a transfer matrix $X \in \mathcal{L}_p$ is defined pointwise on $\mathbb{T}$ as

$$X(e^{i\omega}) = U(e^{i\omega})\Sigma(e^{i\omega})V^*(e^{i\omega}),$$

where $U, V \in \mathcal{L}_\infty$ and $\Sigma \in \mathcal{L}_p$.

For matrix-valued functions $X(z), Y(z)$ defined on $\mathbb{T}$, define

$$\langle X, Y \rangle = \int_{-\pi}^{\pi} \text{tr}\left(X(e^{i\omega})^* Y(e^{i\omega})\right) \frac{d\omega}{2\pi}$$

and the norms

$$\|X\|_1 = \int_{-\pi}^{\pi} \sqrt{X(e^{i\omega})^* X(e^{i\omega})} \, \frac{d\omega}{2\pi}$$

$$\|X\|_2 = \left(\int_{-\pi}^{\pi} \left\|X(e^{i\omega})\right\|_F^2 \, \frac{d\omega}{2\pi}\right)^{1/2},$$

where $\|\cdot\|_F$ is the Frobenius norm.

When a function in $\mathcal{H}_p$ is evaluated on $\mathbb{T}$, it is to be understood as the radial limit $\lim_{r \to 1^+} X(rz)$. The arguments of transfer matrices will often be omitted when they are clear from the context. Equalities and inequalities involving functions evaluated on $\mathbb{T}$ are to be interpreted as holding almost everywhere on $\mathbb{T}$.

III. Problem Formulation

Consider the system in Fig. 1. The source s, source noise m and channel noise n are assumed to be mutually
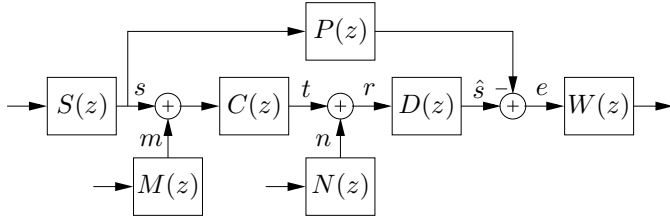
Fig. 2. A representation of the problem in the frequency domain. The transfer functions S, M and N are spectral factors of the source, measurement noise and channel noise, respectively. The delay constraint is determined by P. The encoder and decoder filters are given by C and D. W is an optional frequency weight.

independent, stationary Gaussian[2] sequences with zero mean and known covariance functions. The communication channel has additive noise and a power constraint. That is,

$$r = t + n \tag{5}$$
$$\mathrm{E}(t(k)^2) \leq \sigma^2. \tag{6}$$

Denote the encoder mapping by $\gamma(\cdot)$ and the decoder mapping by $\delta(\cdot)$. The encoder and decoder are assumed to be causal LTI filters with inputs $s+m$ and r, respectively. The estimate of the source sequence is

$$\hat{s} = \delta(t+n) = \delta(\gamma(s+m)+n). \tag{7}$$

Denoting the delay, in number of samples, by d, the reconstruction error is

$$e(k) = s(k-d) - \hat{s}(k). \tag{8}$$

The objective is to choose the encoder and decoder to minimize the stationary value of the MSE, or $\mathrm{E}(e(k)^2)$, subject to the power constraint.

Due to the linearity assumption, the problem can be formulated in the frequency domain, as is illustrated in Fig. 2. In this formulation, all the inputs are mutually independent, zero mean, white noise sequences with unit variance. The transfer functions $S(z), M(z)$ and $N(z)$ are spectral factors of the sequences s, m and n, respectively. The encoder and the decoder are represented by the transfer functions $C(z)$ and $D(z)$. In this formulation, the problem has been generalized in two aspects:

- The delay is replaced by a general LTI filter P. That is, the objective is to estimate the source signal after it has passed through P.
- The error e is passed through a LTI filter W, representing a frequency weighting function, before minimization.

It is assumed that $S, M, N, P, W \in \mathcal{H}_\infty$, that N, W are invertible in $\mathcal{H}_\infty$ and that

$$\exists \varepsilon > 0 \text{ such that } SS^* + MM^* \geq \varepsilon \text{ on } \mathbb{T}, \tag{9}$$

[2]Since only linear solutions are considered, it does not matter if the source, measurement noise or the channel noise are Gaussian or not. Linear solutions may, of course, be more or less suboptimal depending on the distributions.

which implies that S and M have no common zeros on the unit circle (an equivalent condition if $S(z)$ and $M(z)$ are rational functions).

The objective is to choose C and D to minimize the stationary variance of e after filtering by W. By expressing the z-transform of e in terms of the transfer functions in Fig. 2, this quantity can be expressed as

$$J(C,D) = \|W(P-DC)S\|_2^2 + \|WDCM\|_2^2 + \|WDN\|_2^2. \tag{10}$$

Similarly, the power constraint on t can be written as

$$\|CS\|_2^2 + \|CM\|_2^2 \leq \sigma^2. \tag{11}$$

It follows from (10) and (11) that C and D need to be square integrable on the unit circle in order for $J(C,D)$ to be finite and the power constraint to be satisfied. Since the encoder and decoder also should be causal and stable this implies that the optimization should be performed over $C, D \in \mathcal{H}_2$.

IV. Optimal Linear Encoder and Decoder

The problem of finding an optimal linear encoder and decoder will first be solved in the scalar case. The solution will then, under some additional assumptions, be generalized to the vector case.

A. Scalar case

The objective function $J(C,D)$ is clearly not convex in the pair (C,D) due to the appearance of the product DC. In order to find a minimum, the optimization problem will be solved in two steps.

The idea is to first consider the product DC as given and then to find an optimal factorization of this product. The factorization gives an analytical expression for the cost in terms of the product, which means that optimization of the objective may then be performed over the product. When an optimal product is found, the optimality conditions from the solution to the factorization problem can then be applied to find optimal C and D.

First, however, it will be shown that the power constraint (11) can be equivalently written as

$$\|CH\|_2^2 \leq \sigma^2, \tag{12}$$

where the function H has some nice properties.

Lemma 1: Suppose that $S, M \in \mathcal{H}_\infty$ and that (9) holds. Then there exists $H \in \mathcal{H}_\infty$ with $H^{-1} \in \mathcal{H}_\infty$ such that

$$HH^* = SS^* + MM^* \text{ on } \mathbb{T}. \tag{13}$$

Proof: By (9) and the factorization theorem in [27] there exists an outer function $H \in \mathcal{H}_2$ such that (13) holds. Since $S, M \in \mathcal{H}_\infty$ it follows that $H \in \mathcal{H}_\infty$. Moreover, it follows from (9) that $\|H^{-1}\|_\infty \leq 1/\sqrt{\varepsilon}$ and since H is outer it then follows from Lemma 4 (in the appendix) that $H^{-1} \in \mathcal{H}_\infty$. ∎

Now, introduce $K = DC \in \mathcal{H}_1$. The objective (10) can then be written as

$$\|W(P-K)S\|_2^2 + \|WKM\|_2^2 + \|WDN\|_2^2. \tag{14}$$

Note that the first two terms are constant for fixed K. The minimum over C and D, given K, is thus obtained by minimizing the third term in (14) subject to (12) and $K = DC$. This minimization problem is called the optimal factorization problem.

The interpretation is that for any given product of the encoder and decoder, the contribution to the objective of the signals that pass through both the encoder and the decoder is not affected by the choice of the factors C and D — only their product matters. The channel noise, however, only passes through the decoder, which means that D (and implicitly C since $C = D^{-1}K$) should be chosen to minimize the impact of the channel noise on the objective. The solution to the scalar version of the optimal factorization problem is given by the following lemma.

Lemma 2 (Optimal factorization, scalar case): Suppose that $\sigma > 0$, $K \in \mathcal{H}_1$ and that $H, N, W \in \mathcal{H}_\infty$ are invertible in $\mathcal{H}_\infty$. Then the optimization problem

$$\underset{C,D \in \mathcal{H}_2}{\text{minimize}} \|WDN\|_2^2 \qquad (15)$$

subject to

$$K = DC, \quad \|CH\|_2^2 \le \sigma^2 \qquad (16)$$

attains the minimum value

$$\frac{1}{\sigma^2} \|WKHN\|_1^2 . \qquad (17)$$

Moreover, if K is not identically zero then $C, D \in \mathcal{H}_2$ are optimal if and only if $DC = K$ and

$$|C|^2 = \frac{\sigma^2}{\|WKHN\|_1} \left| \frac{WKN}{H} \right| \text{ on } \mathbb{T}. \qquad (18)$$

If $K = 0$, then the minimum is achieved by $D = 0$ and any function $C \in \mathcal{H}_2$ that satisfies $\|CH\|_2^2 \le \sigma^2$.

Proof: If $K = 0$ the proof is trivial, so assume that K is not identically zero. Then C is not identically zero and $D = KC^{-1}$. Then (16) and Cauchy-Schwarz's inequality gives

$$\|WDN\|_2^2 = \|WKC^{-1}N\|_2^2 \ge \frac{\|CH\|_2^2}{\sigma^2} \|WKC^{-1}N\|_2^2$$

$$\ge \frac{1}{\sigma^2} \left\langle |CH|, |WKC^{-1}N| \right\rangle^2 = \frac{1}{\sigma^2} \|WKHN\|_1^2$$

This shows that (17) is a lower bound on the value. Equality holds if and only if $|WKC^{-1}N|$ and $|CH|$ are proportional on $\mathbb{T}$ and $\|CH\|_2^2 = \sigma^2$. It is easily verified that this is equivalent to (18). Thus, C and D achieve the lower bound if and only if $D = KC^{-1}$ and (18) holds.

It remains to show existence of such $C, D \in \mathcal{H}_2$. Note that $WKNH^{-1} \in \mathcal{H}_1$ is not identically zero. Hence, by Theorem 17.17 in [25], $\log |WKNH^{-1}| \in \mathcal{L}_1$. It follows from the factorization theorem in [27] that there exists an outer $C \in \mathcal{H}_2$ that satisfies (18). Thus

$$\|KC^{-1}\|_2^2 = \frac{1}{\sigma^2} \|WKHN\|_1 \|W^{-1}KHN^{-1}\|_1 < \infty,$$

so $D = KC^{-1} \in \mathcal{L}_2$. Since $K \in \mathcal{H}_1$ and $C \in \mathcal{H}_2$ is outer it follows from Lemma 4 (in the appendix) that $D = KC^{-1} \in \mathcal{H}_2$. ∎

Remark 1: Optimal D satisfy

$$|D|^2 = \frac{\|WKHN\|_1}{\sigma^2} \left| \frac{KH}{WN} \right| \text{ on } \mathbb{T}. \qquad (19)$$

Apparently, the magnitudes of C and D are both proportional to the square root of the magnitude of K. This provides some intuition to why the minimum value depends on the 1-norm of K.

Remark 2: The existence part of Lemma 2 shows that a particular solution, where C is outer, can be obtained. By using the freedom available in spectral factorization, it is possible to obtain other solutions, for example by changing the sign of both C and D, or by instead choosing D to be outer. More generally, in the rational case, any non-minimum phase zeros or time delays could be located in C or D.

For any given K an optimal encoder-decoder pair, under the constraint that their product is K, is specified by (18) and (19), respectively. An optimal K can in turn be obtained by inserting the minimum value of $\|WDN\|_2^2$ into (14) and minimizing

$$\varphi(K) = \|W(P - K)S\|_2^2 + \|WKM\|_2^2$$
$$+ \frac{1}{\sigma^2} \|WK [S \quad M] N\|_1^2$$

over K. This is a convex problem. That this procedure in fact solves the main problem is shown by the following theorem, which is the main result of this paper.

Theorem 1: Suppose that $M, N, S, P, W \in \mathcal{H}_\infty$, where N and W are invertible in $\mathcal{H}_\infty$, that $\sigma > 0$ and that (9) holds. Then the optimization problem

$$\underset{C,D \in \mathcal{H}_2}{\text{minimize}} J(C, D) \qquad (20)$$

subject to

$$\|CS\|_2^2 + \|CM\|_2^2 \le \sigma^2 \qquad (21)$$

attains a minimum value that is equal to the minimum of the convex optimization problem

$$\underset{K \in \mathcal{H}_2}{\text{minimize}} \varphi(K), \qquad (22)$$

which is attained by a unique minimizer.

Moreover, suppose $K \in \mathcal{H}_2$ is a solution to (22). If K is not identically zero, then C and D solve (20) subject to (21) if and only if $C \in \mathcal{H}_2$, $D = KC^{-1} \in \mathcal{H}_2$ and

$$|C|^2 = \frac{\sigma^2}{\|WKN [S \quad M]\|_1} \frac{|WKN|}{\sqrt{|S|^2 + |M|^2}} \text{ on } \mathbb{T}. \qquad (23)$$

If $K = 0$, then the solution to (20) and (21) is given by $D = 0$ and any function $C \in \mathcal{H}_2$ that satisfies (21).

Proof: Define $H \in \mathcal{H}_\infty$ according to Lemma 1. Then (21) is equivalent to $\|CH\|_2^2 \le \sigma^2$. Define the sets

$$\Theta = \left\{ (C, D) : C, D \in \mathcal{H}_2, \ \|CH\|_2^2 \le \sigma^2 \right\}$$
$$\Theta(K) = \{(C, D) : (C, D) \in \Theta, \ K = DC \} .$$

Then the infimum of $J(C, D)$ subject to (21) can be written

$$
\begin{aligned}
&\inf_{C, D \in \Theta} J(C, D) \\
&= \inf_{K \in \mathcal{H}_1} \inf_{C, D \in \Theta(K)} J(C, D) \\
&= \inf_{K \in \mathcal{H}_1} \left(\|W(P - K)S\|_2^2 + \|WKM\|_2^2 + \inf_{C, D \in \Theta(K)} \|WDN\|_2^2 \right) \\
&= \inf_{K \in \mathcal{H}_1} \|W(P - K)S\|_2^2 + \|WKM\|_2^2 + \frac{1}{\sigma^2} \|WKHN\|_1^2 \\
&= \inf_{K \in \mathcal{H}_1} \varphi(K) \quad\quad (24)
\end{aligned}
$$

The first equality is true by Theorem 17.10 in [25]. The second equality follows because the first two terms in $\inf_{C, D \in \Theta(K)} J(C, D)$ are constant. The third equality follows from application of Lemma 2 to perform the inner minimization. The final equality follows from (13).

It will now be shown that the minimum is attained in (24) by a unique $K \in \mathcal{H}_2$. Completion of squares gives that

$$
\begin{aligned}
\varphi(K) &= \|W(P - K)S\|_2^2 + \|WKM\|_2^2 + \frac{1}{\sigma^2} \|WKHN\|_1^2 \\
&= \|WPS\|_2^2 + \|WKH\|_2^2 \\
&\quad - 2 \operatorname{Re}\langle WPSS^*, WKHH^{-1} \rangle + \frac{1}{\sigma^2} \|WKHN\|_1^2 \\
&= \left\| WPSS^* H^{-*} - WKH \right\|_2^2 + \frac{1}{\sigma^2} \|WKHN\|_1^2 + \eta,
\end{aligned}
$$

where η is a constant that does not depend on K. Let $X = WKH$ and $R = WPSS^* H^{-*} \in \mathcal{L}_\infty$. Minimizing $\varphi(K)$ over $K \in \mathcal{H}_1$ is then equivalent to minimizing

$$
\psi(X) = \|R - X\|_2^2 + \frac{1}{\sigma^2} \|XN\|_1^2 \quad\quad (25)
$$

over $X \in \mathcal{H}_1$. In the latter problem, it is sufficient to consider X such that $\psi(X) \leq \psi(0) = \|R\|_2^2$. That is, only X satisfying

$$
\begin{aligned}
\|X\|_2 &= \|R - X - R\|_2 \leq \|R - X\|_2 + \|R\|_2 \\
&\leq \sqrt{\psi(X)} + \|R\|_2 \leq 2 \|R\|_2 \overset{\text{def}}{=} r.
\end{aligned}
$$

Now, in the weak topology, $\psi(X)$ is lower semicontinuous on $\mathcal{L}_2$ and the set $\{X : \|X\|_2 \leq r\}$ is compact. This proves the existence of a minimum. The minimum is unique since $\psi(X)$ is strictly convex. Moreover, since $\|X\|_2 \leq r$, it is sufficient to minimize over $X \in \mathcal{H}_2$ instead of $\mathcal{H}_1$.

Suppose now that $X \in \mathcal{H}_2$ minimizes $\psi(X)$. From $H^{-1}, W^{-1} \in \mathcal{H}_\infty$ it follows that $K = W^{-1} X H^{-1} \in \mathcal{H}_2$ attains the infimum value in (24) and that this value is equal to the minimum of (22). Since the minimum is attained in (24) and, by Lemma 2, there exists $(C, D) \in \Theta$ such that $J(C, D) = \varphi(K)$, it follows that the minimum of (20) subject to (21) is attained.

The optimality condition (23) follows from the application of Lemma 2, using that $|H| = \sqrt{|S|^2 + |M|^2}$. ∎

Remark 3: $\varphi(K)$ is convex, and $\varphi(\overline{K}) = \varphi(K)$. Thus,

$$
\varphi\left(\frac{K + \overline{K}}{2}\right) \leq \frac{1}{2}\left(\varphi(K) + \varphi(\overline{K})\right) = \varphi(K).
$$

Since the optimal K is unique, this shows that the minimizing K satisfies $K(e^{-i\omega}) = \overline{K(e^{i\omega})}$. Thus, C can be chosen to have this property as well, meaning that C can be approximated by a rational function with real coefficients. The same holds for D.

Remark 4: It was noted in Remark 2 that the optimal factorization problem can have multiple solutions. To clarify, the optimal K is unique but there are multiple factorizations of K into C and D that achieve the minimum value of $J(C, D)$.

It is noted that the solution of the problem essentially amounts to minimizing the sum of a 2-norm and a 1-norm of the decision variable. The 2-norm represents the cost in the Wiener filter problem, and the 1-norm represents the contribution of the channel noise to the error variance. The SNR σ^2 determines the relative importance of the two terms. For small SNR, the optimal K will have small magnitude since the channel noise dominates the transmitted signal. As the SNR becomes larger, the magnitude of K will become larger, and it will approach the Wiener filter in the limit when the SNR goes to infinity.

B. Vector case

In this section, the results in the previous section will be generalized to the case of vector-valued signals.

Consider again the system in Figure 2 and assume that all signals are vector-valued and all systems are given by their corresponding transfer matrices. The number of elements in signal s is denoted n_s and so forth. That is, $s(k) \in \mathbb{R}^{n_s}$ Matrix dimensions are not explicitly stated in this section except when necessary. It is generally assumed that all matrices are of appropriate size. In addition to all the assumptions made in the scalar case, it is now also assumed that:

1) The communication channel consists of n_t parallel AWGN channels. The power constraint (6) is replaced by the total power constraint

$$
\mathrm{E}(t(k)^T t(k)) \leq \sigma^2.
$$

2) All input signals in Fig. 2 have identity covariance matrices. Moreover, $N(z) = W(z) = I$. That is, the channel noise is white with identity covariance and the frequency weight is uniform.

3) The number of elements in the signals satisfy

$$
n_t \geq \min\{n_s, n_e\}, \quad\quad (26)
$$

where C is $n_t \times n_s$ and D is $n_e \times n_t$. If the number of channels n_t would be smaller than n_f and n_e, then the product DC could not have full rank. This means that optimization over $K = DC$ would have to include a rank constraint, which is very difficult to handle even in the static case.

4) The inequality (9) is replaced by the matrix version

$$
\exists \varepsilon > 0 \text{ such that } FF^* + GG^* \succeq \varepsilon I \text{ on } \mathbb{T}. \quad (27)
$$

The objective is thus to minimize

$$
J_v(C, D) = \|(P - DC)S\|_2^2 + \|DCM\|_2^2 + \|D\|_2^2
$$

subject to

$$\|CS\|_2^2 + \|CM\|_2^2 \leq \sigma^2 \tag{28}$$

The objective and the constraint are thus quite similar to the ones in the scalar case. It will be seen that the equivalent convex problem looks the same but that the optimality condition will, however, be more complicated. The main difference between the scalar and vector versions of the problem is that the optimal factorization (Lemma 2) is much more difficult to prove in the vector case.

Lemma 3 (Optimal factorization, vector case):
Suppose that $\sigma > 0$, $K \in \mathcal{H}_1$, that $H \in \mathcal{H}_\infty$ is invertible in $\mathcal{H}_\infty$ and that (26) holds. Then the optimization problem

$$\underset{C,D \in \mathcal{H}_2}{\text{minimize}} \|D\|_2^2$$

subject to

$$K = DC, \quad \|CH\|_2^2 \leq \sigma^2$$

attains the minimum value $\frac{1}{\sigma^2}\|KH\|_1^2$.

Moreover, suppose that K is not identically zero and let $K = K_i K_o$ be an inner-outer factorization and $K_o H = U_o \Sigma V^*$ be a singular value decomposition. Then $C, D \in \mathcal{H}_2$ are optimal if and only if

$$K = DC, \quad \|CH\|_2^2 = \sigma^2, \tag{29}$$

$$DD^* = \frac{\|KH\|_1}{\sigma^2} K_i U_o \Sigma U_o^* K_i^*. \tag{30}$$

If $K = 0$ then the minimum is achieved by $D = 0$ and any function $C \in \mathcal{H}_2$ that satisfies $\|CH\|_2^2 \leq \sigma^2$.

Proof: If $K = 0$ the proof is trivial, so assume that K is not identically zero. Then neither C nor D are identically zero and $\alpha = \|CH\|_2 > 0$. Now, suppose that C, D are feasible and that $\alpha < \sigma$. Then

$$C_\alpha = \frac{\sigma}{\alpha} C, \quad D_\alpha = \frac{\alpha}{\sigma} D$$

are feasible and $\|D_\alpha\|_2 < \|D\|_2$. Hence, a necessary condition for optimality is that $\|CH\|_2^2 = \sigma^2$.

The remainder of this proof is divided into three parts. First, the dual problem is considered. Then, it is shown that there is a saddle point and the optimality criteria are derived. Finally, existence of the solution is proven by construction.

Dual Problem: In order to avoid dealing with analyticity constraints associated with $\mathcal{H}_2$, the search will temporarily be relaxed to $C, D \in \mathcal{L}_2$. Later, it will be shown that there are $C, D \in \mathcal{H}_2$ that satisfy the derived optimality criteria. For $\lambda \geq 0$ and matrix-valued $\Phi \in \mathcal{L}_\infty$, introduce the Lagrangian

$$L(C, D, \lambda, \Phi) = \|D\|_2^2 + \lambda\left(\|CH\|_2^2 - \sigma^2\right)$$
$$- \langle \operatorname{Re}\Phi, \operatorname{Re}DC - K\rangle - \langle \operatorname{Im}\Phi, \operatorname{Im}DC - K\rangle$$
$$= \|D\|_2^2 + \lambda\left(\|CH\|_2^2 - \sigma^2\right) - \operatorname{Re}\langle \Phi, DC - K\rangle$$
$$= \int_{-\pi}^{\pi} \|D\|_F^2 + \lambda\|CH\|_F^2 - \operatorname{Re}\operatorname{tr}\left(\Phi^*(DC - K)\right)\frac{d\omega}{2\pi} - \lambda\sigma^2 \tag{31}$$

The integrand in (31) can be rewritten as

$$\|D\|_F^2 + \lambda\|CH\|_F^2 - \operatorname{Re}\operatorname{tr}\left(C\Phi^*D - \Phi^*K\right)$$
$$= \left\|D - \frac{1}{2}\Phi C^*\right\|_F^2 + \lambda\|CH\|_F^2 - \frac{1}{4}\|C\Phi^*\|_F^2 + \operatorname{Re}\operatorname{tr}\left(\Phi^*K\right)$$
$$= \left\|D - \frac{1}{2}\Phi C^*\right\|_F^2 + \operatorname{tr}\left[C\left(\lambda HH^* - \frac{1}{4}\Phi^*\Phi\right)C^* + \operatorname{Re}\Phi^*K\right] \tag{32}$$

Only the first term depends on D. The contribution of this term is minimized by

$$D = \frac{1}{2}\Phi C^*. \tag{33}$$

If (33) holds, then L only depends on C through the first term inside the brackets in (32). Pointwise minimization of that term gives

$$\inf_{C \in \mathcal{L}_2}\ \operatorname{tr}\left[C\left(\lambda HH^* - \frac{1}{4}\Phi^*\Phi\right)C^*\right]$$
$$= \begin{cases} 0, & 4\lambda HH^* \geq \Phi^*\Phi \text{ on } \mathbb{T} \\ -\infty, & \text{otherwise.} \end{cases}$$

Moreover, the remaining term in (32) can be written

$$\operatorname{tr}\left(\Phi^*K\right) = \operatorname{tr}\left(\Phi^*DC\right) = \frac{1}{2}\operatorname{tr}\left(C\Phi^*\Phi C^*\right) = \frac{1}{2}\|\Phi C^*\|_F^2.$$

Thus, $\operatorname{tr}\left(\Phi^*K\right)$ is real and non-negative, and

$$\inf_{C,D \in \mathcal{L}_2} L = \begin{cases} \int_{-\pi}^{\pi}\operatorname{tr}\left(\Phi^*K\right)\frac{d\omega}{2\pi} - \lambda\sigma^2, & 4\lambda HH^* \geq \Phi^*\Phi \text{ on } \mathbb{T} \\ -\infty, & \text{otherwise.} \end{cases}$$

Introduce

$$\Psi = \frac{1}{2\sqrt{\lambda}}\Phi H^{-*}.$$

Then the dual problem can be written as

$$\underset{\lambda \geq 0, \Psi \in \mathcal{L}_\infty}{\text{maximize}}\ 2\sqrt{\lambda}\int_{-\pi}^{\pi}\operatorname{tr}\left(\Psi^*KH\right)\frac{d\omega}{2\pi} - \lambda\sigma^2$$

subject to

$$\Psi^*\Psi \leq I \text{ on } \mathbb{T}. \tag{34}$$

The dual function is concave in λ. Letting $\lambda = 0$ gives the value 0. Since $\operatorname{tr}\left(\Psi^*KH\right) \geq 0$ there exists $\lambda > 0$ that gives a positive value, so the optimal λ is given by the first-order condition

$$\left(\frac{1}{\sigma^2}\int_{-\pi}^{\pi}\operatorname{tr}\left(\Psi^*KH\right)\frac{d\omega}{2\pi}\right)^2 = \lambda,$$

obtained by differentiation with respect to λ. With this λ the dual problem simplifies to

$$\underset{\Psi \in \mathcal{L}_\infty}{\text{maximize}}\ \frac{1}{\sigma^2}\left(\int_{-\pi}^{\pi}\operatorname{tr}\left(\Psi^*KH\right)\frac{d\omega}{2\pi}\right)^2 \tag{35}$$

subject to (34).

The integrand in (35) will now be maximized pointwise. Recall that $KH = K_i K_o H = K_i U_o \Sigma V^*$ and denote the number of rows of K_o by m. Then Σ is diagonal with

diagonal elements σ_k, $k = 1 \ldots m$. Since K is $n_e \times n_s$ the rank of K is not greater than $\min\{n_e, n_s\}$ and thus

$$m \leq \min\{n_e, n_f\}. \tag{36}$$

K_o is row outer by definition and H is outer by Corollary 4.7 in [24]. It follows that $K_o H$ is row outer and thus has full row rank. It follows that the singular values are positive: $\sigma_k > 0$, $k = 1 \ldots m$. Since $K_o H$ is wide (it has $n_s \geq m$ columns) it follows that U_o is square and thus unitary.

Define $U = K_i U_o$ and $\widetilde{\Psi} = U^* \Psi V$. Then it follows from (34) and $UU^* \leq I$ that

$$\widetilde{\Psi}^* \widetilde{\Psi} = V^* \Psi^* UU^* \Psi V \leq V^* \Psi^* \Psi V \leq V^* V = I.$$

Using $\widetilde{\Psi}$, an upper bound can be obtained for the integrand in (35):

$$\begin{aligned} \sup_{\Psi^* \Psi \leq I} \mathrm{tr}\,(\Psi^* K H) &= \sup_{\Psi^* \Psi \leq I} \mathrm{tr}\,(\Psi^* U \Sigma V^*) \\ &= \sup_{\Psi^* \Psi \leq I} \mathrm{tr}\,(V^* \Psi^* U \Sigma) \\ &\leq \sup_{\widetilde{\Psi}^* \widetilde{\Psi} \leq I} \mathrm{tr}\,\left(\widetilde{\Psi}^* \Sigma\right) \\ &= \sum_{k=1}^m \sup_{|\widetilde{\Psi}_{kk}| \leq 1} \sigma_k \widetilde{\Psi}_{kk} = \sum_{k=1}^m \sigma_k \end{aligned}$$

The supremum is achieved if and only if $\widetilde{\Psi} = I$. Therefore, the upper bound is achieved by Ψ if and only if $U^* \Psi V = I$ and $\Psi^* \Psi \leq I$. The set of Ψ satisfying these conditions can be parametrized as:

$$\Psi = UV^* + \Psi_0 = K_i U_o V^* + \Psi_0 \tag{37}$$
$$I \geq \Psi^* \Psi, \tag{38}$$

where Ψ_0 satisfies

$$0 = U^* \Psi_0 V = U_o^* K_i^* \Psi_0 V. \tag{39}$$

Pre-multiplying (39) with U_o gives the equivalent condition

$$K_i^* \Psi_0 V = 0. \tag{40}$$

Choosing, for example, $\Psi_0 = 0$ gives $\Psi = UV^*$, which attains the upper bound. Hence, the value of the dual problem is

$$\begin{aligned} \max_{\Psi^* \Psi \leq I} &\frac{1}{\sigma^2} \left(\int_{-\pi}^{\pi} \mathrm{tr}\,(\Psi^* K H) \frac{d\omega}{2\pi} \right)^2 \\ &= \frac{1}{\sigma^2} \left(\int_{-\pi}^{\pi} \mathrm{tr}\,(VU^* U \Sigma V^*) \frac{d\omega}{2\pi} \right)^2 = \frac{1}{\sigma^2} \|KH\|_1^2. \end{aligned}$$

The maximizing dual variables are given by

$$\Phi = 2\sqrt{\lambda} \Psi H^* = 2\sqrt{\lambda}(K_i U_o V^* + \Psi_0) H^* \tag{41}$$

where Ψ_0 is such that (37), (38) and (40) hold, and

$$\lambda = \left(\frac{1}{\sigma^2} \|KH\|_1 \right)^2. \tag{42}$$

Saddle Point: It will now be shown that there is a saddle point, which implies that the duality gap is zero.

In the following, assume that (37), (38), (40), (41) and (42) hold. Then λ and Φ are dual feasible. The point (C, D, λ, Φ) is a saddle point if and only if $C, D \in \mathcal{H}_2$ are primal feasible,

$$\lambda \left(\|CH\|_2^2 - \sigma^2 \right) = 0 \tag{43}$$

and

$$L(C, D, \lambda, \Phi) = \inf_{\widehat{C}, \widehat{D} \in \mathcal{H}_2} L(\widehat{C}, \widehat{D}, \lambda, \Phi). \tag{44}$$

The saddle point conditions imply that $\|CH\|_2 = \sigma$ since $\lambda > 0$ and that $D = \frac{1}{2} \Phi C^*$ as it was seen earlier that this follows from minimization of the Lagrangian.

Suppose that the saddle point conditions hold. Then C, D satisfy $K = DC$ and $D = \frac{1}{2} \Phi C^*$. Moreover,

$$\begin{aligned} DD^* &= \frac{1}{2} DC\Phi^* = \frac{1}{2} K \Phi^* = \sqrt{\lambda} K_i K_o H(V U_o^* K_i^* + \Psi_0^*) \\ &= \sqrt{\lambda}(K_i U_o \Sigma U_o^* K_i^* + K_i U_o \Sigma V^* \Psi_0^*). \end{aligned}$$

Clearly, DD^* and $K_i U_o \Sigma U_o^* K_i^*$ are Hermitian. Accordingly,

$$A = K_i U_o \Sigma V^* \Psi_0^*$$

must be Hermitian. Now, by (40),

$$AK_i = K_i U_o \Sigma V^* \Psi_0^* K_i = 0$$
$$\Rightarrow 0 = AK_i = A^* K_i = \Psi_0 V \Sigma U_o^* K_i^* K_i = \Psi_0 V \Sigma U_o^*.$$

Hence, $A = 0$ and

$$DD^* = \sqrt{\lambda} K_i U_o \Sigma U_o^* K_i^* = \frac{\|KH\|_1}{\sigma^2} K_i U_o \Sigma U_o^* K_i^*. \tag{45}$$

Suppose instead that $C, D \in \mathcal{H}_2$ satisfy $K = DC$, $\|CH\|_2 = \sigma$ and (45). Then C, D are primal feasible and (43) is satisfied. Moreover,

$$\begin{aligned} L(C, D, \lambda, \Phi) &= \|D\|_2^2 = \sqrt{\lambda} \int_{-\pi}^{\pi} \mathrm{tr}\,(K_i U_o \Sigma U_o^* K_i^*) \frac{d\omega}{2\pi} \\ &= \sqrt{\lambda} \int_{-\pi}^{\pi} \mathrm{tr}\,(\Sigma) \frac{d\omega}{2\pi} = \frac{1}{\sigma^2} \|KH\|_1^2, \end{aligned}$$

so (44) holds and thus the saddle point conditions are satisfied. Since these assumptions and the saddle point conditions imply each other, they are equivalent.

To conclude, it has been shown that (C, D, λ, Φ) is a saddle point, which implies that $C, D \in \mathcal{H}_2$ achieve the claimed minimum, if and only if $K = DC$, $\|CH\|_2^2 = \sigma^2$ and (45) holds.

Existence of Solution: Define $B = \sqrt{\lambda} U_o \Sigma U_o^* \in \mathcal{L}_1$, which is Hermitian with real diagonal. Recall that $K_o H$ is row outer with singular values $\sigma_k > 0$, $k = 1 \ldots m$. From this and Lemma 5 it follows that $\log \sigma_k \in \mathcal{L}_1$. Since U_o is unitary it also follows that B is positive definite. Moreover,

$$\log \det B = \frac{m}{2} \log \lambda + \sum_{k=1}^m \log \sigma_k \in \mathcal{L}_1$$

Therefore, according to the theorem in [27], there is an outer transfer matrix $D_o \in \mathcal{H}_2$ such that $B = D_o D_o^*$. Let $\widetilde{D} = K_i D_o \in \mathcal{H}_2$ and $\widetilde{C} = D_o^{-1} K_o$. Then

$$\widetilde{C} = D_o^{-1} K_o H H^{-1} = D_o^{-1} U_o \Sigma V^* H^{-1}$$
$$= D_o^{-1} U_o \Sigma U_o^* U_o V^* H^{-1} = \frac{1}{\sqrt{\lambda}} D_o^* U_o V^* H^{-1} \in \mathcal{L}_2$$

Since D_o is outer it follows from Lemma 4 that $\widetilde{C} \in \mathcal{H}_2$.

It can now be verified that $\widetilde{C}$ and $\widetilde{D}$ satisfy the optimality conditions:

$$\widetilde{D}\widetilde{C} = K_i D_o D_o^{-1} K_o = K_i K_o = K,$$

$$\left\| \widetilde{C} H \right\|_2^2 = \left\| D_o^{-1} K_o H \right\|_2^2 = \int_{-\pi}^{\pi} \operatorname{tr} \left(H^* K_o^* D_o^{-*} D_o^{-1} K_o H \right) \frac{d\omega}{2\pi}$$
$$= \int_{-\pi}^{\pi} \operatorname{tr} \left(V \Sigma U_o^* B^{-1} U_o \Sigma V^* \right) \frac{d\omega}{2\pi}$$
$$= \frac{1}{\sqrt{\lambda}} \int_{-\pi}^{\pi} \operatorname{tr} \left(\Sigma \right) \frac{d\omega}{2\pi} = \sigma^2$$

and

$$\widetilde{D}\widetilde{D}^* = K_i D_o D_o^* K_i^* = \sqrt{\lambda} K_i U_o \Sigma U_o^* K_i^*.$$

If the rank of K does not equal n_t, then $\widetilde{C}$ and $\widetilde{D}$ are not of the required dimensions. $\widetilde{C}$ is $m \times n_s$ and $\widetilde{D}$ is $n_e \times m$, where, by (26) and (36), $m \leq \min\{n_e, n_f\} \leq n_t$. It is required that C is $n_t \times n_s$ and that D is $n_e \times n_t$. To solve this problem, let

$$D = \begin{bmatrix} \widetilde{D} & 0_{n_e \times n_t - m} \end{bmatrix} \in \mathcal{H}_2, \quad C = \begin{bmatrix} \widetilde{C} \\ 0_{n_t - m \times n_s} \end{bmatrix} \in \mathcal{H}_2.$$

Noting that $DC = \widetilde{D}\widetilde{C} = K$, that $\|CH\|_2 = \left\| \widetilde{C} H \right\|_2$ and that $DD^* = \widetilde{D}\widetilde{D}^*$ it is finally concluded that C, D are optimal. ∎

Just as in the scalar case, the solution to the optimal factorization problem can be used to find an equivalent convex problem. This problem looks exactly the same both cases. The theorem for the vector case is now stated.

Theorem 2: Suppose that $\sigma > 0$, $S, M, P \in \mathcal{H}_\infty$ and that (26) and (27) hold. Then the optimization problem

$$\underset{C,D \in \mathcal{H}_2}{\text{minimize}} \ J(C, D) \tag{46}$$

subject to

$$\|CS\|_2^2 + \|CM\|_2^2 \leq \sigma^2 \tag{47}$$

attains a minimum value that is equal to the minimum of the convex optimization problem

$$\underset{K \in \mathcal{H}_2}{\text{minimize}} \ \|(P - K)S\|_2^2 + \|KM\|_2^2 + \frac{1}{\sigma^2} \left\| K \begin{bmatrix} S & M \end{bmatrix} \right\|_1^2, \tag{48}$$

which is attained by a unique minimizer.

Moreover, suppose $K \in \mathcal{H}_2$ is a solution to (48). If K is not identically zero, then $C, D \in \mathcal{H}_2$ solve (46) subject to (47) if and only if

$$K = DC, \quad \left\| C \begin{bmatrix} S & M \end{bmatrix} \right\|_2^2 = \sigma^2,$$
$$DD^* = \frac{\left\| K \begin{bmatrix} S & M \end{bmatrix} \right\|_1}{\sigma^2} K_i U_o \Sigma U_o^* K_i^*,$$

where K_i is defined by an inner-outer factorization $K = K_i K_o$ and U_o and Σ are given by a singular value decomposition $K_o H = U_o \Sigma V^*$, where $H \in \mathcal{H}_\infty$ satisfies $H^{-1} \in \mathcal{H}_\infty$ and $HH^* = SS^* + MM^*$.

If $K = 0$, then the solution to (46) and (47) is given by $D = 0$ and any function $C \in \mathcal{H}_2$ that satisfies (47).

Proof: With the assumption (27), Lemma 1 holds in the matrix case as well. The rest of the proof is identical to the proof of Theorem 1, except that Lemma 3 is used instead of Lemma 2, with the obvious implications for the optimality conditions. ∎

Remark 5: The assumption (26) may deserve some explanation. If there are too few communication channels relative to the dimensionality of s and e, the maximum rank of the product DC may be smaller than the smallest dimension of K. Then not all K would be realizable as a product of D and C, and a rank condition would have to be imposed on K in Theorem 2. In principle, this changes nothing, but the assumption is included in order to avoid formulating the solution in terms of an optimization problems that cannot be reliably solved.

C. Optimal LTI Filters Require Infinite Memory

The structure of optimal linear encoders and decoders will now be studied. In particular, it will be shown that the optimal filters generally have non-rational transfer functions. This corresponds to systems with infinite memory, since it is generally impossible to find a finite dimensional state-space realization of such transfer functions.

We consider the scalar case with white channel noise and rational S, M, P and W. This implies that $N(z) = 1$ and that $R = WPSS^*H^{-*}$ is rational, where H satisfies (13). Since S, M, P and W are proper, it can safely be assumed that R is proper. If R is not proper then it can be made proper by multiplying H with z^{-k}, for a large enough k.

If we define

$$\psi(X) = \|R - X\|_2^2 + \frac{1}{\sigma^2} \|X\|_1^2, \tag{49}$$

the solution is given by solving the problem

$$\underset{X \in \mathcal{H}_2}{\text{minimize}} \ \psi(X). \tag{50}$$

Recall that the minimum of (50) is attained and that it is a strictly convex problem. It will now be shown that a necessary condition for the minimum cannot be satisfied by a rational X except in some special cases. To begin with, two simple observations are made:

1) If the solution X to (50) is a rational function it can be factorized into inner-outer factors as $X = FX_o$. The outer factor X_o is then a rational function that solves the optimization problem

$$\underset{X_o \in \mathcal{H}_2}{\text{minimize}} \ \|F^*R - X_o\|_2^2 + \frac{1}{\sigma^2} \|X_o\|_1^2 \tag{51}$$

where F^*R is a rational function. Thus, we can assume without loss of generality that the optimal solution X is outer.

2) Due to the orthogonality

$$\|R - X\|_2^2 = \|R_-\|_2 + \|R_+ - X\|_2^2, \qquad (52)$$

where $R = R_+ + R_-$ is a decomposition of R into the analytical and anti-analytical parts, respectively, we can also assume that R is analytical since the anti-analytical part does not affect the optimization. That is, $R = R_+$.

Another assumption we make to simplify the proof is that the function R has only simple poles. Note that the poles of $P_+(F^*R)$ are the same as of P_+R, so the simplicity of the poles remains true through the two rewritings above.

Theorem 3: Consider the problem (50) with a proper non-constant rational function $R \in \mathcal{H}_2$ and assume that the poles of R are simple and that the optimal solution X is not identically zero. Then X is not a rational function.

Proof: We split the proof into several steps to underline the structure.

Step 1: Calculate the first variation of the functional ψ and state the Euler-Lagrange equation. The standard differentiation of $\psi(X + \epsilon h)$ with respect to ϵ and then setting $\epsilon = 0$ gives

$$\delta\psi(h) = 2\,\mathrm{Re}\int_{-\pi}^{\pi}\left(\frac{\|X\|_1}{\sigma^2}\frac{X}{|X|} - (R - X)\right)^* h\,\frac{d\omega}{2\pi}$$

For convex problems the necessary and sufficient condition for the minimum is that $\delta\psi(h) = 0$ for all $h \in \mathcal{H}_2$. It gives the Euler-Lagrange equation for the optimal X as

$$\frac{\|X\|_1}{\sigma^2}P_+\frac{X}{|X|} = R - X$$

where P_+ is the standard orthogonal projection from $\mathcal{L}_2$ to $\mathcal{H}_2$. Note that the constant $\|X\|_1\,\sigma^{-2}$ can be incorporated into X and R without affecting their rationality. So in the following we assume without loss of generality that this constant is equal to 1 and analyze the equation

$$P_+\frac{X}{|X|} = R - X. \qquad (53)$$

Since X is not identically zero, it is not zero almost everywhere on $\mathbb{T}$ and the fraction $\frac{X}{|X|}$ is well defined.

Step 2: We will now assume that the solution $X \in \mathcal{H}_2$ to (53) is rational and show that it will lead us to a contradiction. In this step we prove that rationality of R and X in (53) implies rationality of $|X|$. Indeed

$$|X| = X^*\frac{X}{|X|} = X^*P_+\frac{X}{|X|} + X^*P_-\frac{X}{|X|}.$$

The second term in the right hand side is anti-analytical, hence

$$P_+|X| = P_+\left(X^*P_+\frac{X}{|X|}\right). \qquad (54)$$

Clearly $P_+X|X|^{-1}$ is rational due to (53) and thus the right hand side of (54) is rational too. Accordingly, $P_+|X|$ is also rational. Furthermore, the function $|X|$ is real and has a symmetric Laurent series. Therefore, the function $|X|$ must be rational itself.

Factorization as $|X| = h^*h = |h|^2 = |h^2|$ with an outer rational $h \in \mathcal{H}_2$ and assuming, as was explained previously, that X is outer, gives the only possibility that $X = h^2$. That is, the rational solution X must be a square of a rational function.

Step 3: Rewrite the Euler-Lagrange equation in terms of h and then in terms of numerators and denominators of h and R. Substituting $X = h^2$ into (53) gives

$$P_+\frac{X}{|X|} = P_+\frac{h^2}{h^*h} = P_+\frac{h}{h^*} = R - h^2.$$

Introduce the notations for the numerators and the denominators

$$h = \frac{p}{q}, \quad R = \frac{b}{a}$$

where a, b, p and q are polynomials. The polynomials a, p and q are stable by definition, since $R \in \mathcal{H}_2$ and h is outer in $\mathcal{H}_2$. Introduce the notation for the conjugate polynomial to p as

$$\tilde{p}(z) = z^n p(z^{-1})$$

where n is the degree of p. The conjugate of a stable polynomial has the same degree and is anti-stable. With these notations in mind the Euler-Lagrange equation becomes

$$P_+\frac{p\tilde{q}z^{n-m}}{q\tilde{p}} = \frac{b}{a} - \frac{p^2}{q^2} = \frac{bq^2 - ap^2}{aq^2}. \qquad (55)$$

Here n and m are degrees of p and q respectively.

Step 4: Calculate the projection in the left hand side of (55) and state the polynomial version of the Euler-Lagrange equation. We assume now that $n - m \geq 0$ and cover the opposite in the next step. Perform the partial fraction decomposition

$$\frac{p\tilde{q}z^{n-m}}{q\tilde{p}} = \frac{Q}{q} + \frac{r}{\tilde{p}}$$

where Q is a polynomial and the degree of r is less than n. Then

$$P_+\frac{p\tilde{q}z^{n-m}}{q\tilde{p}} = \frac{Q}{q}$$

and the equation (55) becomes

$$aqQ = bq^2 - ap^2.$$

Clearly $q(z) = 0$ implies $a(z) = 0$ since p and q are prime, hence $a = qa_0$ where a_0 is a polynomial. Canceling q above we get

$$a_0qQ = bq - a_0p^2.$$

Similarly $q(z) = 0$ implies $a_0(z) = 0$ and thus $a_0 = qa_1$. Canceling again gives

$$a_1qQ = b - a_1p^2.$$

Now it is clear that $a_1 = 1$ since otherwise $a_1(z) = 0$ would give $b(z) = 0$, which is impossible since a and b are also prime. Finally, we have $a = q^2$, which contradicts the assumption that zeros of a are simple unless $q = a = 1$. But for a proper non-constant R it is impossible.

Step 5: The case $n - m < 0$ is similar. Denote $k = m - n$. The only difference is in the partial fraction decomposition

$$\frac{p\tilde{q}}{q\tilde{p}z^k} = \frac{Q}{q} + \frac{r}{\tilde{p}z^k}$$

where Q is a polynomial and the degree of r is less than $n + k = m$. The rest is exactly the same as in Step 4 with the same conclusion that $a = q^2$ which contradicts the assumption. ∎

Because S, M and W are assumed to be rational and $X = WKH$ it follows that K is rational if and only if X is rational. Clearly, if K is not rational, it cannot be factorized as $K = DC$ with rational C and D. Thus, the transfer functions of optimal LTI encoders and decoders are not rational.

As explained previously, this means that the filters can not be realized using finite memory. Obviously, approximations have to be done for practical implementation. For example, impulse responses of the filters may be truncated. It remains to investigate the impact on the performance of such approximations.

If the channel has noise-free feedback, that is, if C has access to the channel output, then C can estimate the states of D exactly. It would be interesting to study if the memory of optimal linear encoders and decoders could be bounded in this case. Such a result would also be in line with the structural result for causal coders in [11], where the memory was bounded given that the encoder has knowledge of the decoder state.

V. Numerical Solution

A procedure for obtaining an approximate numerical solution will now be outlined for the vector version of the problem.

1) The first step is to solve the optimization problem (48) or, alternatively, minimize (25) (the constant part η must then be added to obtain the distortion). An approximate solution can be obtained by using a finite basis representation of K and approximating the integrals by sums over a finite number of frequency grid points. Such an approximated problem can be cast as a quadratic program with second-order cone constraints.
2) Perform a matrix spectral factorization of $SS^* + MM^*$ to obtain $H \in \mathcal{H}_\infty$ with $H^{-1} \in \mathcal{H}_\infty$.
3) Perform an inner-outer factorization of K to obtain $K_i K_o = K$.
4) Perform a singular value decomposition of $K_o H$ to obtain $U_o \Sigma V^* = K_o H$.
5) Use a finite basis approximation $A(\omega)$ of DD^*, for example using the parametrization

$$A(\omega) = A_0 + \sum_{k=1}^{N_c} A_k \left(e^{ki\omega} + e^{-ki\omega}\right)$$

and fit $A(\omega)$ to

$$\frac{\left\| K \begin{bmatrix} S & M \end{bmatrix} \right\|_1}{\sigma^2} K_i U_o \Sigma U_o^* K_i^*,$$

by minimizing the deviation in some suitable norm.
6) Perform a spectral factorization of $A(\omega)$, choosing D_o as the stable and outer spectral factor.
7) Let $D = K_i D_o$ and $C = D_o^{-1} K_o$.
8) If C and D are of incorrect size, add rows of zeros to C and columns of zeros to D until they are of correct size.

In the scalar case, the procedure is simplified as follows: Step 2 and 6 requires only scalar spectral factorizations, step 3, 4 and 8 are skipped and step 5 consists of fitting $A(\omega)$ to

$$\frac{\left\| K \begin{bmatrix} S & M \end{bmatrix} \right\|_1}{\sigma^2} |KH|.$$

A. Example

The numerical solution is illustrated by the following example. Consider the problem with $S = 1/(z - 0.9)$, $M = 0$, $W = N = 1$ and $P = z^{-d}$. The functional $\psi(X)$, given by (25), was approximated by discretization of the integrals over 4000 grid points, uniformly placed on the unit circle. X was parametrized as an FIR filter with 60 coefficients. The minimization was then carried out for different SNR levels σ^2 and delays d, using Matlab, Yalmip [28] and SeDuMi [29].

The resulting MSE distortion levels are displayed in Fig. 3 together with the OPTA for the case with no delay constraint, obtained from (1). It can be seen that for small SNR's, the distortion is very close to the lower bound. This is not surprising since for zero SNR, the minimum distortion is $\|WPS\|_2^2 = \|S\|_2^2$ over any type of coding system. For medium SNR's, the distortion is lower for longer delays. The difference seems, however, to decrease when the SNR becomes larger. The gap to the OPTA seems to approach about a factor two for high SNR's, regardless of delay. This suggests that for this source, it is the linearity, rather than the delay constraint, that is the performance-limiting factor for high SNR levels.

VI. Conclusion

This paper has shown how to find optimal LTI encoders and decoders for joint source-channel coding for Gaussian sources and channels. It has also been shown that such encoders and decoders in general require infinite memory. Thus, some approximation has to be done for numerical solution of the problem. It would be interesting to investigate if the performance loss due to such approximations can be somehow bounded.

In the scalar case, the solution has been extended to handle channels with feedback [1]. This is not presented here to conserve space. Another extension is the problem of feedback control over AWGN channels, which will be the topic of an upcoming paper.

Possible topics for further research includes extending the solution in the MIMO case to channels with colored noise, investigating memory bounds when the channel has feedback and the suboptimality of linear solutions.
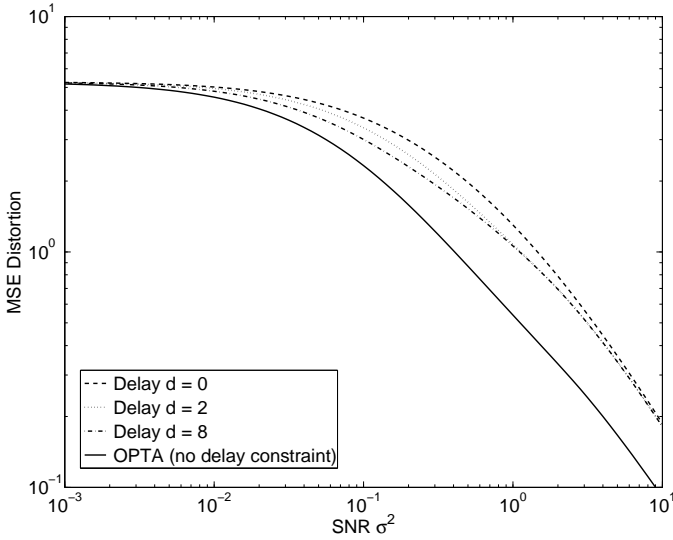
Fig. 3. MSE distortion as a function of SNR level (logarithmic scale) for optimal linear encoders and decoders for three different delay constraints (approximate solutions), and the OPTA for the case without delay constraint. (Problem parameters: $S = 1/(z-0.9)$, $M = 0$, $W = N = 1$, $P = z^{-d}$)

Appendix

Lemma 4: Suppose $Y \in \mathcal{N}^+$ is square and outer, $X \in \mathcal{N}^+$, and that $Y^{-1}X \in \mathcal{L}_p$. Then $Y^{-1}X \in \mathcal{H}_p$.

Proof: $Y^{-1} \in \mathcal{N}^+$ by Theorem 10 in [26]. It is easy to verify that the product of two $\mathcal{N}^+$ functions is $\mathcal{N}^+$. The proof follows from the fact that $\mathcal{L}_p \cap \mathcal{N}^+ = \mathcal{H}_p$ [24]. ∎

Lemma 5: Suppose that $m \leq n$ and that the $m \times n$ transfer matrix $X \in \mathcal{H}_p$, $p \in \{1, 2, \infty\}$, is row outer. Then the singular values of X satisfy

$$\log \sigma_k \in \mathcal{L}_1, \quad k = 1 \ldots m.$$

Proof: By Theorem 8 in [26] there exists a factorization $X = X_{co}X_i$, where X_{co} is column outer and X_i is inner. Since X_{co} has full column rank on $\mathbb{T}$ it cannot have more columns than rows, and since X is row outer X_{co} cannot have fewer rows than columns. Thus X_{co} is $m \times m$ and hence, by Theorem 10 in [26], $\det X_{co}$ is outer and thus $\det X_{co} \in \mathcal{N}^+$. According to a statement in section 17.19 in [25] it follows that $\log |\det X_{co}| \in \mathcal{L}_1$.

For the singular values of X, it holds that

$$\sum_{k=1}^{m} \log \sigma_k = \frac{1}{2} \log \prod_{k=1}^{m} \sigma_k^2 = \frac{1}{2} \log \det XX^*$$

$$= \frac{1}{2} \log \det X_{co}X_iX_i^*X_{co}^* = \frac{1}{2} \log \det X_{co}X_{co}^*$$

$$= \log |\det X_{co}| \in \mathcal{L}_1.$$

Furthermore, $\sigma_k \in \mathcal{L}_1$ since $X \in \mathcal{H}_p$. Because $\log \sigma_k < \sigma_k$ it holds that

$$\int_{-\pi}^{\pi} \log \sigma_k \, d\omega < \int_{-\pi}^{\pi} \sigma_k \, d\omega < \infty, \quad k = 1 \ldots m$$

Since the sum of the logarithms is $\mathcal{L}_1$ and every term has an integral bounded from above, it follows that the integral of every term also must be bounded from below. That is,

$$\int_{-\pi}^{\pi} \log \sigma_k \, d\omega > -\infty, \quad k = 1 \ldots m$$

and hence $\log \sigma_k \in \mathcal{L}_1$, $k = 1 \ldots m$ ∎

Acknowledgment

References

[1] E. Johannesson, "Control and communication with signal-to-noise ratio constraints," Ph.D. dissertation, Department of Automatic Control, Lund University, Sweden, Oct. 2011.

[2] R. J. Pilc, "The optimum linear modulator for a gaussian source used with a gaussian channel," Bell System Technical Journal, vol. 48, no. 9, Nov. 1969.

[3] K.-H. Lee and D. P. Petersen, "Optimal linear coding for vector channels," IEEE Transactions on Communications, vol. 24, no. 12, pp. 1283 – 1290, Dec. 1976.

[4] T. Basar, B. Sankur, and H. Abut, "Performance bounds and optimal linear coding for discrete-time multichannel communication systems (corresp.)," IEEE Transactions on Information Theory, vol. 26, no. 2, pp. 212 – 217, Mar. 1980.

[5] Y. Kochman and R. Zamir, "Analog matching of colored sources to colored channels," IEEE Transactions on Information Theory, vol. 57, no. 6, pp. 3180 –3195, June 2011.

[6] R. Dobrushin and B. Tsybakov, "Information transmission with additional noise," IRE Transactions on Information Theory, vol. 8, no. 5, pp. 293–304, Sept. 1962.

[7] J. Wolf and J. Ziv, "Transmission of noisy information to a noisy receiver with minimum distortion," IEEE Transactions on Information Theory, vol. 16, no. 4, pp. 406–411, July 1970.

[8] T. Berger, Rate distortion theory: a mathematical basis for data compression. Englewood Cliffs, NJ, USA: Prentice-Hall, 1971.

[9] H. Witsenhausen, "Indirect rate distortion problems," IEEE Transactions on Information Theory, vol. 26, no. 5, pp. 518–521, Sept. 1980.

[10] D. Neuhoff and R. Gilbert, "Causal source codes," IEEE Transactions on Information Theory, vol. 28, no. 5, pp. 701–713, Sept. 1982.

[11] H. Witsenhausen, "On the structure of real-time source coders," Bell Syst. Tech. J., vol. 58, no. 6, pp. 1437–1451, July-Aug 1979.

[12] J. Walrand and P. Varaiya, "Optimal causal coding–decoding problems," IEEE Transactions on Information Theory, vol. 29, no. 6, pp. 814–820, Nov. 1983.

[13] D. Teneketzis, "On the structure of optimal real-time encoders and decoders in noisy communication," Information Theory, IEEE Transactions on, vol. 52, no. 9, pp. 4017 –4035, sept. 2006.

[14] M. Gastpar, "To code or not to code," Ph.D. dissertation, EPFL, Lausanne, 2002.

[15] M. S. Derpich and J. Østergaard, "Improved upper bounds to the causal quadratic rate-distortion function for gaussian stationary sources," IEEE Transactions on Information Theory, accepted for publication.

[16] M. Derpich, E. Silva, D. Quevedo, and G. Goodwin, "On optimal perfect reconstruction feedback quantizers," IEEE Transactions on Signal Processing, vol. 56, no. 8, pp. 3871–3890, Aug. 2008.

[17] M. Derpich, "Optimal source coding with signal transfer function constraints," Ph.D. dissertation, University of Newcastle, 2009.

[18] V. Borkar, S. Mitter, and S. Tatikonda, "Optimal sequential vector quantization of markov sources," in Proc. IEEE Conference on Decision and Control, vol. 1, 2001, pp. 205–210.

[19] S. Yüksel, "On optimal causal coding of partially observed markov sources in single and multi-terminal settings," 2010. [Online]. Available: http://arxiv.org/abs/1010.4824v2

[20] A. Mahajan and D. Teneketzis, "On the design of globally optimal communication strategies for real-time noisy communication systems with noisy feedback," IEEE Journal on Selected Areas in Communications, vol. 26, no. 4, pp. 580–595, May 2008.

[21] ——, "Optimal design of sequential real-time communication systems," IEEE Transactions on Information Theory, vol. 55, no. 11, pp. 5317–5338, Nov. 2009.

[22] N. Wiener, Extrapolation, Interpolation, and Smoothing of Stationary Time Series. The MIT Press, 1964.

[23] N. Martins, M. Dahleh, and J. Doyle, "Fundamental limitations of disturbance attenuation in the presence of side information," IEEE Transactions on Automatic Control, vol. 52, no. 1, pp. 56–66, Jan. 2007.

[24] J. Garnett, Bounded analytic functions, revised 1st ed. New York, NY, USA: Springer, 2007.

[25] W. Rudin, Real and Complex Analysis, 3rd ed. McGraw-Hill Science/Engineering/Math, May 1986.

[26] Y. Inouye, "Linear systems with transfer functions of bounded type: Canonical factorization," IEEE Transactions on Circuits and Systems, vol. 33, no. 6, pp. 581–589, June 1986.

[27] N. Wiener and E. Akutowicz, "A factorization of positive Hermitian matrices," Indiana Univ. Math. J., vol. 8, pp. 111–120, 1959.

[28] J. Löfberg, "Yalmip : A toolbox for modeling and optimization in MATLAB," in Proceedings of the CACSD Conference, Taipei, Taiwan, 2004.

[29] J. Sturm, "Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones," Optimization Methods and Software, vol. 11–12, pp. 625–653, 1999, version 1.3 available from `http://sedumi.ie.lehigh.edu/`.